

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR LETTERS PATENT

Semi-Automatic Annotation of Multimedia Objects

Inventor(s):

Wen-Yin Liu
Hong-Jiang Zhang

ATTORNEY'S DOCKET NO. MS1-605US

070968-10000

000007 8820750

1 **TECHNICAL FIELD**

2 This invention relates to systems and methods for annotating multimedia
3 objects, such as digital images, to facilitate keyword-based retrieval methods.
4

5 **BACKGROUND**

6 The popularity of digital images is rapidly increasing due to improving
7 digital imaging technologies and convenient availability facilitated by the Internet.
8 More and more digital images are becoming available every day. The images are
9 kept in image databases, and retrieval systems provide an efficient mechanism for
10 users to navigate through the growing numbers of images in the image databases.

11 Traditional image retrieval systems allow users to retrieve images in one of
12 two ways: (1) keyword-based image retrieval or (2) content-based image retrieval.
13 Keyword-based image retrieval finds images by matching keywords from a user
14 query to keywords that have been added to the images. Content-based image
15 retrieval (CBIR) finds images that have low-level image features similar to those
16 of an example image, such as color histogram, texture, shape, and so forth.
17 However, CBIR has a drawback in that searches may return entirely irrelevant
18 images that just happen to possess similar features. Since content-based image
19 retrieval has a low performance level, keyword-based image search is more
20 preferable.

21 To facilitate keyword-based image retrieval, the images (or generally,
22 multimedia objects) must first be labeled with one or more keywords. Labeling
23 semantic content of images, or multimedia objects, with a set of keywords is a
24 process known as image (or multimedia) annotation. Annotated images can be
25 found using keyword-based search, while un-annotated image cannot.

000001-8820250

1 semantically relevant keywords. The annotation process is performed in
2 background, hidden from the user, while the user conducts searches.

3 The annotation process is "semi-automatic" in that it utilizes both keyword-
4 based information retrieval and content-based image retrieval techniques to
5 automatically search for multimedia objects, and then encourages users to provide
6 feedback on the retrieved objects. The user is asked to identify the returned
7 objects as either relevant or irrelevant to the query keywords and based on this
8 feedback, the system automatically annotates the objects with semantically
9 relevant keywords and/or updates associations between the keywords and objects.

10 In the described implementation, the system performs both keyword-based
11 and content-based retrieval. A user interface allows a user to specify a query in
12 terms of keywords and/or examples objects. Depending on the input query, the
13 system finds multimedia objects with keywords that match the keywords in the
14 query and/or objects with similar content features. The system ranks the objects
15 and returns them to the user.

16 The user interface allows the user to identify multimedia objects that are
17 more relevant to the query, as well as objects that are less or not relevant. The
18 system monitors the user feedback using a combination of feature-based relevance
19 feedback and semantic-based relevance feedback.

20 If the multimedia object is deemed relevant by the user and is not yet
21 annotated with the keyword, the system adds the keyword to the object. The
22 objects and keywords are maintained in a database and a semantic network is
23 constructed on top of the database to define associations between the keywords
24 and objects. Weights are assigned to the keyword-object associations to indicate
25 how relevant the keyword is to the object.

During the retrieval-feedback-annotation cycle, the system adjusts the weights according to the user feedback, thereby strengthening associations between keywords and objects identified as more relevant and weakening the associations between keywords and objects identified as less relevant. If the association becomes sufficiently weak, the system removes the keyword from the multimedia object.

Accordingly, the semi-automatic annotation process captures the efficiency of automatic annotation and the accuracy of manual annotation. As the retrieval-feedback-annotation cycle is repeated, both annotation coverage and annotation quality of the object database is improved.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of an exemplary computer network in which a server computer implements a multimedia object retrieval/annotation system that may be accessed over a network by one or more client computers.

Fig. 2 is a block diagram of the retrieval/annotation system architecture.

Fig. 3 illustrates a first screen view of a user interface for the retrieval/annotation system.

Fig. 4 illustrates a semantic network that represents relationships between keywords and multimedia objects.

Fig. 5 illustrates a second screen view of the user interface for the retrieval/annotation system.

Fig. 6 is a flow diagram of an initial query handling process in which a user initially submits a keyword query for a multimedia object.

1 Fig. 7 is a flow diagram of a refinement and annotation process in which
2 the retrieval/annotation system learns from the user's feedback pertaining to how
3 relevant the objects are to the initial query and annotates the objects accordingly.

4 5 **DETAILED DESCRIPTION**

6 This disclosure describes an annotation system for annotating multimedia
7 objects, such as digital images, video clips, and audio objects, with semantically
8 relevant keywords. The annotation system employs a "semi-automatic"
9 annotation technique that captures the efficiency of automatic annotation and the
10 accuracy of manual annotation. The semi-automatic annotation technique
11 employs both keyword-based information retrieval and content-based image
12 retrieval techniques to automate searches for objects, and then encourages users to
13 provide feedback to the result set of objects. The user identifies objects as either
14 relevant or irrelevant to the query keywords and based on this feedback, the
15 system automatically updates associations between the keywords and objects. As
16 the retrieval-feedback-annotation cycle is repeated, the annotation coverage and
17 annotation quality of the object database is improved.

18 The annotation process is accomplished in a hidden/implicit fashion,
19 without the user's notice. As the user naturally uses the multimedia object
20 database, more and more objects are annotated and the annotations become more
21 and more accurate. The result is a set of keywords associated with each individual
22 multimedia object in the database.

23 The annotation system is described in the context of an Internet-based
24 image retrieval system that searches and retrieves images from an image database.
25 It is noted, however, that the invention pertains to other multimedia objects

1 besides digital images. Furthermore, the system may be implemented in other
2 environments, such as a non-networked computer system. For instance, this
3 technology may be applied to stand-alone image database systems.

4 5 Exemplary System

6 Fig. 1 shows an exemplary computer network system 100 that implements
7 an annotation system for annotating multimedia objects, such as digital images,
8 with semantically relevant keywords. In the described implementation, the
9 annotation system is integrated with a retrieval system that searches and retrieves
10 objects from a database using both keyword-based retrieval techniques and
11 content-based retrieval techniques.

12 The network system 100 includes a client computer 102 that submits
13 queries to a server computer 104 via a network 106, such as the Internet. While
14 the system 100 can be implemented using other networks (e.g., a wide area
15 network or local area network) and should not be limited to the Internet, the
16 system will be described in the context of the Internet as one suitable
17 implementation. The web-based system allows multiple users to perform retrieval
18 tasks simultaneously at any given time.

19 The client 102 is representative of many diverse computer systems,
20 including general-purpose computers (e.g., desktop computer, laptop computer,
21 etc.), network appliances (e.g., set-top box (STB), game console, etc.), and the
22 like. The client 102 includes a processor 110, a volatile memory 112 (e.g., RAM),
23 and a non-volatile memory 114 (e.g., ROM, Flash, hard disk, optical, etc.). The
24 client 102 also has one or more input devices 116 (e.g., keyboard, keypad, mouse,
25

1 searches. Through the iterative feedback, annotations are added to the objects in a
2 hidden fashion, thereby continually adapting and improving the semantic network
3 utilized in the keyword-based retrieval. The annotation process yields tremendous
4 advantages in terms of both efficiency and accuracy.

6 Retrieval And Annotation System Architecture

7 Fig. 2 illustrates the retrieval/annotation system architecture 140 in more
8 detail. The UI 150 has a query interface 200 that accepts text-based keyword or
9 natural language queries as well as content-based queries resulting from selection
10 of an example image (or other type of media object).

11 Fig. 3 shows an example of a query interface screen 300 presented by the
12 user interface 150 for entry of a query. The screen 300 presents a natural language
13 text entry area 302 that allows user to enter keywords, phrases, or complete
14 sentences. After entering one or more words, the user actuates a button 304 that
15 initiate the search for relevant objects. Alternatively, the user can browse a pre-
16 defined concept hierarchy by selecting one of the categories listed in section 306
17 of the query screen 300. The user actuates the category link to initiate a search for
18 objects within the category.

19 With reference again to Fig. 2, the query is passed to the query handler 152.
20 In the illustrated implementation, the query handler 152 includes a natural
21 language parser 202 to parse text-based queries, such as keywords, phrases, and
22 sentences. The parser 202 is configured to extract keywords from the query, and
23 may utilize syntactic and semantic information from natural language queries to
24 better understand and identify keywords. The parsed results are used as input to
25 the semantic network that associates keywords with images in the database 142.

09702288-103000

1 closeness in features between two objects. The set of highest-ranked objects are
2 returned to the user interface 200 and presented to the user for consideration.

3 The user interface 150 has an object browser 218 that allows the user to
4 browse the various objects returned from the keyword-based and content-based
5 search. The returned objects are presented in scrollable pages, or as thumbnails in
6 one or more pages.

7 Fig. 5 shows an example results screen 500 containing a set of image
8 objects returned in response to the user entering the keyword "tiger" into the text
9 entry area 302 of query screen 300 (Fig. 3). Depending on display size, one or
10 more images are displayed in the results screen 500. Here, six images 502(1)-
11 502(6) are displayed at one time. If there are more images than can be displayed
12 simultaneously, navigation "Next" and "Prev" buttons 504 are presented to permit
13 browsing to other images in the result set.

14 The user interface allows the user to feedback relevance information as
15 he/she browses the images. Each image has several feedback options. For
16 instance, each image has a "View" link 506 that allows the user to enlarge the
17 image for better viewing. Activation of a "Similar" link 508 initiates a subsequent
18 query for images with both similar semantic content and similar low-level features
19 as the corresponding image. This refined search will be presented in the next
20 screen and this process may be repeated many times until the user finds a set of
21 images that are highly relevant to the query.

22 Furthermore, each image has both positive and negative relevance marks
23 that may be individually selected by the user. The relevance marks allow the user
24 to indicate on an image-by-image basis, which images are more relevant to the
25 search query and which are less relevant. Examples of such marks include a "+"

1 and “-” combination, or a “thumbs up” and “thumbs down”, or a change in
2 background color (e.g., red means less relevant, blue means more relevant).

3 In the example screen 500, images 502(1), 502(2), and 502(5) are marked
4 with a blue background, indicating a positive match that these images do in fact
5 represent tigers. Images 502(4) and 502(6) have a red background, indicating that
6 the do not match the query “tiger”. Notice closely that these images contain
7 leopards and not tigers. Finally, image 502(3) has a gradient background (neither
8 positive nor negative) and will not be considered in the relevance feedback. This
9 image presents a wolf, which has essentially no relevance to tigers.

10 After providing relevant feedback, the user activates the “Feedback” button
11 510 to submit the feedback to the feedback analyzer 156. The learning begins at
12 this point to improve the image retrieval process for future queries.

13 Turning again to Fig. 2, the feedback analyzer 156 monitors this user
14 feedback. A relevance feedback monitor 220 tracks the feedback and performs
15 both semantic-based relevance feedback and low-level feature relevance feedback
16 in an integrated fashion. The feedback analyzer 156 further implements a machine
17 learning algorithm 222 to train the semantic-based retrieval model and the feature-
18 based retrieval model based on the relevance feedback to thereby improve the
19 results for future search efforts on the same or similar keywords. One particular
20 implementation of an integrated framework for semantic-based relevance feedback
21 and feature-based relevance feedback is described below in more detail under the
22 heading “Integrated Relevance Feedback Framework”.

23 The annotator 158 uses the relevance feedback to annotate the objects in the
24 database 142. In this manner, annotation takes place in a hidden way whenever
25 relevance feedback is performed. The annotator 158 assigns initial keywords to

1 the objects in response to user queries, thereby creating the links in the semantic
2 network 400. The annotator 158 continually adjust the weights assigned to
3 keyword-object links over time as the user continues the search and refinement
4 process.

5 The retrieval/annotation system 140 offers many advantages over
6 conventional systems. First, it locates images using both keywords and low-level
7 features, thereby integrating keyword-based image retrieval and content-based
8 image retrieval. Additionally, it integrates both semantic-based relevance
9 feedback and feature-based relevance feedback. A further benefit is the semi-
10 automatic annotation process that takes place in the background. As the query-
11 retrieval-feedback process iterates, the system annotates objects and modifies the
12 semantics network.

13 Retrieval and Annotation Process

14
15 Figs. 6 and 7 show a retrieval and annotation process implemented by the
16 system 140 of Fig. 2. The process entails a first phase for producing an original
17 object result set from an initial query (Fig. 6) and a second phase for refining the
18 search efforts, training the search models and annotating the objects based on user
19 feedback to the result set (Fig. 7). In one implementation, the image retrieval
20 process is implemented as computer executable instructions that, when executed,
21 perform the operations illustrated as blocks in Figs. 6 and 7.

22 For discussion purposes, the process is described in the context of an image
23 retrieval system for retrieving images from the image database. However, the
24 process may be implemented using other types of multimedia objects. The
25 process further assumes that a coarse concept hierarchy of the available images

000001-88220750

1 exists. For instance, images of people may be coarsely annotated generally as
2 "people" and more particularly as "men" and "women". In addition, the low-level
3 features of the images in the image database 142 may be calculated offline and
4 correlated with the images through a data structure. This removes any potential
5 slowdown caused by computing low-level features during the image retrieval
6 process.

7 At block 602, the retrieval/annotation system 140 receives an initial query
8 submitted by a user via the user interface 150. Suppose the user enters a search
9 query to locate images of "tigers" by, for example, entering any of the following
10 queries into the query screen 300 (Fig. 3):

11
12 "tigers"

13 "tiger pictures"

14 "Find pictures of tigers"

15 "I'm looking for images of tigers."
16

17 At block 604, the query handler 152 parses the user query to extract one or
18 more keywords. In our example, the keyword "tiger" can be extracted from
19 anyone of the queries. Other words, such as "pictures" and "images" may also be
20 extracted, but we'll focus on the keyword "tiger" for illustration purposes.

21 At block 606, the retrieval/annotation system 140 automatically searches
22 the image database 142 to identify images annotated with the keyword "tiger".
23 The system may also simultaneously search of similar words (e.g., cat, animal,
24 etc.). Block 606 distinguishes between two possible situations. In the first case,
25 there are some images already annotated with the keyword(s) that match the

000001-8820760

1 keyword, the annotator 158 annotates that image in the image database with the
2 keywords from the query and assigns an initial weight to the association link in the
3 semantic network. As an example, the initial link might be assigned a weight
4 value of "1". If the image has already been annotated, the weight of this keyword
5 for this image is increased with some given increment, such as "1", so that over
6 time, the weight of strongly associated keywords and images grows large. A large
7 weight represents a higher confidence that the search is accurate when keywords
8 are used to identify images.

9 The annotator 158 also adjusts the annotations for irrelevant images and/or
10 modifies the weighting of the semantic network (block 708). For each irrelevant
11 image, the weight of the keyword-image link is decreased by some value. In one
12 implementation, the weight is reduced by one-fourth of its original value. If the
13 weight becomes very small (e.g., less than 1), the annotator 158 removes the
14 keyword from the annotation of this image. It is noted that there may be many
15 methods that can be used to re-weight the keywords during the annotation process,
16 and the above re-weighting scheme is only an exemplary implementation.

17 At block 710, the retrieval/annotation process performs another retrieval
18 cycle based on the user feedback to refine the search. The results are once again
19 presented to the user for analysis as to their relevancy.

20 Block 712 accounts for the situation where the original query did not return
21 any relevant images, nor did the user find an example image to refine the search.
22 In this situation, the retrieval/annotation system simply outputs images in the
23 database one page at a time to let the user browse through and select the relevant
24 images to feed back into the system.

25

Integrated Relevance Feedback Framework

This section described on exemplary implementation of integrating semantic-based relevance feedback with low-level feature-based relevance feedback. Semantic-based relevance feedback can be performed relatively easily compared to its low-level feature counterpart. One exemplary implementation of semantic-based relevance feedback is described first, followed by how this feedback can be integrated with feature-based relevance feedback.

For semantic-based relevance feedback, a voting scheme is used to update the weights w_{ij} associated with each link in the semantic network 300 (Fig. 3). The weight updating process is described below.

Step 1: Initialize all weights w_{ij} to 1. That is, every keyword is initially given the same importance.

Step 2: Collect the user query and the positive and negative feedback examples.

Step 3: For each keyword in the input query, check if any of them is not in the keyword database. If so, add the keyword(s) into the database without creating any links.

Step 4: For each positive example, check if any query keyword is not linked to it. If so, create a link with weight "1" from each missing keyword to this image. For all other keywords that are already linked to this image, increase the weight by "1".

Step 5: For each negative example, check to see if any query keyword is linked with it. If so, set the new weight $w_{ij} = w_{ij}/4$. If the weight w_{ij} on any link is less than 1, delete that link.

1
2 It can be easily seen that as more queries are input, the system is able to
3 expand its vocabulary. Also, through this voting process, the keywords that
4 represent the actual semantic content of each image are assigned larger weights.

5 As noted previously, the weight w_{ij} associated on each keyword-image link
6 represents the degree of relevance in which this keyword describes the linked
7 image's semantic content. For retrieval purposes, another consideration is to
8 avoid having certain keywords associated with a large number of images in the
9 database. The keywords with many links to many images should be penalized.
10 Therefore, a relevance factor r_{ij} of the i^{th} keyword association to the j^{th} image be
11 computed as follows:

12
13
$$r_{ij} = w_{ij} (\log_2 \frac{M}{d_i} + 1)$$

14

15
16 where M is the total number of images in the database, and d_i is the number of
17 links that the i^{th} keyword has.

18 Now, the above semantic-based relevance feedback needs to be integrated
19 with the feature-based relevance feedback. It is known from previous research
20 (See, Rui, Y., Huang, T. S. "A Novel Relevance Feedback Technique in Image
21 Retrieval," ACM Multimedia, 1999) that the ideal query vector q_i^* for feature i is
22 the weighted average of the training samples for feature i given by:

23
$$q_i^{r*} = \frac{\pi^T X_i}{\sum_{n=1}^N \pi_n} \quad (3)$$

24
25

where X_i is the $N \times K_i$ training sample matrix for feature i , obtained by stacking the N training vectors x_{ni} into a matrix, and where N is an element vector $\pi = [\pi_1, \dots, \pi_N]$ that represents the degree of relevance for each of the N input training samples. The optimal weight matrix W_i^* is given by:

$$W_i^* = (\det(C_i))^{\frac{1}{K_i}} C_i^{-1} \quad (4)$$

where C_i is the weighted covariance matrix of X_i . That is:

$$C_{i,rs} = \frac{\sum_{n=1}^N \pi_n (x_{nir} - q_{ir})(x_{nis} - q_{is})}{\sum_{n=1}^N \pi_n} \quad r, s = 1, K \quad K_i \quad (5)$$

The critical inputs into the system are x_{ni} and π . Initially, the user inputs these data to the system. However, this first step can be eliminated by automatically providing the system with this initial data. This is done by searching the semantic network for keywords that appear in the input query. From these keywords, the system follows the links to obtain the set of training images (duplicate images are removed). The vectors x_{ni} can be computed easily from the training set. The degree of relevance vector π is computed as follows:

$$\pi_i = \alpha^M \sum_{j=1}^M r_{ij} \quad (6)$$

where M is the number of query keywords linked to the training image i , r_{ij} is the relevance factor of the i^{th} keyword associated with image j , and $\alpha > 1$ is a suitable constant. The degree of relevance of the j^{th} image increases exponentially with the number of query keywords linked to it. In the one implementation, an experimentally determined setting of $\alpha = 2.5$ yielded the best results.

To incorporate the low-level feature based feedback and ranking results into high-level semantic feedback and ranking, a unified distance metric function G_j is defined to measure the relevance of any image j within the image database in terms of both semantic and low-level feature content. The function G_j is defined using a modified form of the Rocchio's formula as follows:

$$G_j = \log(1 + \pi_j) D_j + \beta \left\{ \frac{1}{N_R} \sum_{k \in N_R} \left[\left(1 + \frac{I_1}{A_1} \right) S_{jk} \right] \right\} - \gamma \left\{ \frac{1}{N_N} \sum_{k \in N_N} \left[\left(1 + \frac{I_2}{A_2} \right) S_{jk} \right] \right\} \quad (7)$$

where D_j is the distance score computed by the low-level feedback, N_R and N_N are the number of positive and negative feedbacks respectively, I_1 is the number of distinct keywords in common between the image j and all the positive feedback images, I_2 is the number of distinct keywords in common between the image j and all the negative feedback images, A_1 and A_2 are the total number of distinct keywords associated with all the positive and negative feedback images respectively, and finally S_{ij} is the Euclidean distance of the low-level features between the images i and j .

The first parameter α in Rocchio's formula is replaced with the logarithm of the degree of relevance of the j^{th} image. The other two parameters β and γ can

1 be assigned a value of 1.0 for simplicity. However, other values can be given to
2 emphasize the weighting difference between the last two terms.

3 Using the method described above, the combined relevance feedback is
4 provided as follows.

5
6 Step 1: Collect the user query keywords

7 Step 2: Use the above method to compute x_{ni} and π and input them into the
8 low-level feature relevance feedback component to obtain the
9 initial query results.

10 Step 3: Collect positive and negative feedbacks from the user.

11 Step 4: Update the weighting in the semantic network according to the 5-
12 step process described earlier in this section.

13 Step 5: Update the weights of the low-level feature based component.

14 Step 6: Compute the new x_{ni} and π and input into the low-level feedback
15 component. The values of x_{ni} may be computed beforehand in a
16 pre-processing step.

17 Step 7: Compute the ranking score for each image using equation 7 and sort
18 the results.

19 Step 8: Show new results and go to step 3.

20
21 The image retrieval system is advantageous over prior art systems in that it
22 learns from the user's feedback both semantically and in a feature based manner.
23 In addition, if no semantic information is available, the process degenerates into
24 conventional feature-based relevance feedback, such as that described by Rui and
25

1 Huang in the above-cited "A Novel Relevance Feedback Technique in Image
2 Retrieval".

3 4 New Object Registration

5 Adding new multimedia objects into the database is a very common
6 operation under many circumstances. For retrieval systems that entirely rely on
7 low-level content features, adding new objects simply involves extracting various
8 feature vectors for the set of new objects. However, since the retrieval system
9 utilizes keywords to represent the objects' semantic contents, the semantic
10 contents of the new objects have to be labeled either manually or automatically.
11 In this section, an automatic labeling technique is described.

12 The automatic labeling technique involves guessing the semantic content of
13 new objects using low-level features. The following is an exemplary process for
14 digital images:

15
16 Step 1: For each category in the database, compute the representative
17 feature vectors by determining the centroid of all images within
18 this category.

19 Step 2: For each category in the database, find the set of representative
20 keywords by examining the keyword association of each image in
21 this category. The top N keywords with largest weights whose
22 combined weight does not exceed a previously determined
23 threshold τ are selected and added into the list of the representative
24 keywords. The value of the threshold τ is set to 40% of the total
25 weight.

Step 3: For each new image, compare its low-level feature vectors against the representative feature vectors of each category. The images are labeled with the set of representative keywords from the closest matching category with an initial weight of 1.0 on each keyword.

Because the low-level features are not enough to present the images' semantics, some or even all of the automatically labeled keywords will inevitably be inaccurate. However, through user queries and feedbacks, semantically accurate keywords labels will emerge while semantically inaccurate keywords will slowly be eliminated.

Another problem related to automatic labeling of new images is the automatic classification of these images into predefined categories. This problem is addressed by the following process:

Step 1: Put the automatically labeled new images into a special “unknown” category.

Step 2: At regular intervals, check every image in this category to see if any keyword association has received a weight greater than a threshold ξ . If so, extract the top N keywords whose combined weight does not exceed the threshold τ .

Step 3: For each image with extracted keywords, compare the extracted keywords with the list of representative keywords from each category. Assign each image to the closest matching category. If none of the available categories result in a meaningful match, leave this image in the “unknown” category.

1
2 The keyword list comparison function used in step 3 of the above algorithm
3 can take several forms. An ideal function would take into account the semantic
4 relationship of keywords in one list with those of the other list. However, for the
5 sake of simplicity, a quick function only checks for the existence of keywords
6 from the extracted keyword list in the list of representative keywords.
7

8 Conclusion

9 Although the description above uses language that is specific to structural
10 features and/or methodological acts, it is to be understood that the invention
11 defined in the appended claims is not limited to the specific features or acts
12 described. Rather, the specific features and acts are disclosed as exemplary forms
13 of implementing the invention.
14
15
16
17
18
19
20
21
22
23
24
25